DOCUMENT RESUME

ED 313 388 TM 014 143

AUTHOR Engelhard, George, Jr.
TITLE Historical Views of the Concept of Invariance and Measurement Theory in the Behavioral Sciences.
SPONS AGENCY National Academy of Education, Washington, D.C.
PUB DATE Mar 89
NOTE 51p.; An earlier version of this paper was presented at the International Objective Measurement Workshop (5th, Berkeley, CA, March 1989).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Behavioral Sciences; Behavior Theories; Item Analysis; Item Sampling; *Latent Trait Theory; *Measurement; Qualitative Research; Science History
IDENTIFIERS *Invariance Principle; Item Calibration; *Item Invariance; Rasch Model

ABSTRACT

A historical perspective on and substantive review of the concept of invariance are provided. Progress made toward solving measurement problems related to invariance is also assessed. Two major classes of invariant measurement are described: (1) sample-invariant item calibration; and (2) item-invariant measurement of individuals. The work of S. S. Stevens is used to help clarify the concept of invariance. The importance of invariance as a key measurement concept is then illustrated via the measurement theories of E. L. Thorndike, L. L. Thurstone, and G. Rasch. The study methodology uses quotations and original figures to illustrate how these researchers addressed measurement problems related to invariance. A comparison and discussion of these three researchers' theories of measurement are presented in terms of their contributions to the solution of problems related to the concept of invariance. Rasch's research is seen as the means by which the issues raised by the other two researchers were resolved. A case is made for viewing invariance as a fundamental aspect of measurement in the behavioral sciences. Invariance appears to be essential in order to realize the advantages of objective measurement. A 58-item list of references, one table, and five figures are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED313388

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
GEORGE ENGELHARD, JR.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

HISTORICAL VIEWS OF THE CONCEPT OF INVARIANCE AND MEASUREMENT THEORY IN THE BEHAVIORAL SCIENCES

George Engelhard, Jr.

Emory University

Address: Professor George Engelhard, Jr.
Emory University
Division of Educational Studies
210 Fishburne Building
Atlanta, GA 30322

Running head: HISTORICAL VIEWS OF INVARIANCE

[invari - Paper presented at the Fifth International Objective Measurement Workshop, University of California, Berkeley]

March 1989

TM014143

BEST COPY AVAILABLE

Abstract

The purpose of this study is to provide a historical perspective on the concept of invariance. Two major classes of invariant measurement are described -- sample-invariant item calibration and item-invariant measurement of individuals. The work of Stevens is used to help clarify the concept of invariance. The importance of invariance as a key measurement concept is then illustrated with the measurement theories of Thorndike, Thurstone and Rasch. A case is made for viewing invariance as a fundamental aspect of measurement in the behavioral sciences; invariance appears to be essential in order to realize the advantages of objective measurement.

HISTORICAL VIEWS OF INVARIANCE: EVIDENCE FROM THE MEASUREMENT THEORIES OF THORNDIKE, THURSTONE AND RASCH

> The history of science is the history
> of measurement (Cattell, 1893, p. 316)
>
> The scientist is usually looking for
> invariance whether he knows it or not
> (Stevens, 1951, p. 20)

Invariance has been identified as a fundamental aspect of measurement in the behavioral sciences (Andrich, 1988a; Bock & Jones, 1968; Jones, 1960; Stevens, 1951). In essence, the goal of invariant measurement has been succinctly stated by Stevens: "the scientist seeks measures that will stay put while his back is turned" (1951, p. 21). The concept of invariance has implications for both item calibration and the measurement of individuals.

Many of the measurement problems that confront us in psychology and education today, such as those related to invariance, are not new. By taking a historical perspective on these measurement problems, we can increase our understanding of the measurement problems themselves, assess the adequacy of solutions proposed by major measurement theorists and identify promising areas for future research. Progress, and in some cases lack of progress, towards the solution of basic measurement problems can also be meaningfully documented.

During the 20th century, there have been two major research traditions which have guided measurement theorists attempting to quantify various human characteristics, such as abilities, aptitudes and attitudes. One tradition has its roots in the psychometric work of Charles Spearman (1904); this research tradition is focused on the test score and is primarily concerned with measurement error and the decomposition of an observed test score into several components including a "true" score and various error components. This research tradition within mental test theory can be labelled "classical test theory". A second research tradition which has developed in a parallel fashion has its roots in the 19th century work in psychophysics and has continued into present practice through the various forms of latent trait theory or more specifically item response theory (IRT). This second research tradition will be referred to as "scaling theory". The focus of research within this second tradition is on the calibration of both individuals and items onto a latent variable scale. Within these two research traditions, classical test theory and scaling theory, there are several dominant perspectives that have evolved over time. For example, Spearman's research on classical test theory has been extended through generalizability theory (Brennan, 1983; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989), as well as the LISREL models

developed by Karl Joreskog (Joreskog & Sorbom, 1986). This paper examines progress within the second measurement tradition of scaling theory due to the contributions of Thorndike, Thurstone and Rasch; measurement perspectives within classical test theory will not be addressed in detail here.

A great deal of educational and psychological research has been conducted within the framework of classical test theory and empirical research workers routinely include "coefficient alphas" or "KR-20s" for the instruments used in their studies. Along with this concern for "reliability" coefficients, research workers have also worried about the validity of their instruments, although documenting what a test score <u>really</u> represents is rarely resolved in most studies and may ultimately be the most important research question of all. Instead of focusing on measurement problems related to reliability and validity which are the central concepts of classical test theory (Loevinger, 1957), this study focuses on measurement problems related to the concept of invariance which appear clearly within scaling theory; this is <u>not</u> to say that the concepts of reliability or especially validity are unimportant, rather that different research traditions focus on different aspects of the measurement problems encountered in the behavioral sciences. In fact, invariance has important relationships to and implications for issues related to reliability and validity, and is

essential for gaining a clear understanding of certain persistent problems encountered in classical test theory. As pointed out by Jones and Appelbaum (1989), developments in item response theory have led to constructive changes in psychological testing and the "primary advantage of IRT over classical test theory resides in properties of invariance" (p. 24).

Purpose

The purpose of this paper is to provide a historical perspective on the concept of invariance. Several enduring measurement problems related to item calibration and the measurement of individuals can be meaningfully viewed using the concept of invariance. The measurement theories of Thorndike, Thurstone and Rasch are used because they address measurement problems related to the concept of invariance and proposed solutions to these problems. These measurement theorists also share a common research tradition based on scaling theory.

Method

Quotations and original figures when available are used to illustrate how Thorndike, Thurstone and Rasch addressed measurement problems related to invariance. Although there are quantitative aspects to the approaches used to address invariance, it is beyond the scope of this paper to provide detailed derivations of the equations used by each theorist to achieve sample-invariant item

calibration and item-invariant measurement of individuals. These derivations are presented in Engelhard (1984) for measurement issues related to sample-invariant item calibration; a parallel analysis can also be developed for issues related to the item-invariant measurement of individuals, but is not included here.

In the next section of this paper, the concept of invariance is defined and arguments are presented for its importance as a key idea in measurement. A description of the measurement theories of Thorndike, Thurstone and Rasch is presented next; the role of invariance in each of these theories is also examined. Next, a comparison and discussion of these three theories of measurement is presented in terms of their contributions to the solution of problems related to the concept of invariance. The final section includes a summary of the major points of this paper, as well as suggestions for additional research in this area.

THE CONCEPT OF INVARIANCE

Within the behavioral sciences, S. S. Stevens (1951) has presented one of the strongest cases for the general importance of the concept of invariance. In his chapter on "Mathematics, Measurement and Psychophysics", which appeared in the Handbook of Experimental Psychology, Stevens describes the role of this concept in mathematics and physics, and he argued that "many psychological problems are already conceived as the deliberate search for

invariances" (p. 20). In fact, Stevens defined the whole field of science in terms of a quest for invariance and the concomitant generalizability of results. In his words,

> The scientist is usually looking for invariance whether he knows it or not. Whenever he discovers a functional relationship his next question follows naturally: under what conditions does it hold? . . . The quest for invariant relations is essentially the aspiration toward generality, and in psychology, as in physics, the principles that have wide applications are those we prize.
>
> (Stevens, 1951, p. 20)

Applying this view of invariance more specifically to measurement issues, Stevens used the concept of invariance to define his familiar scales of measurement -- nominal, ordinal, interval and ratio scales (Stevens, 1946); in his words,

> Each of the four classes of scales is best characterized by its range of invariance -- by the kinds of transformations that leave the "structure" of the scale undistorted. And the nature of invariance sets limits to the kinds of statistical manipulations that can be legitimately applied to the scaled data. (Stevens, 1951, p. 23)

Influenced by the insightful work of Mosier (1940, 1941), Stevens pointed out the symmetry between the fields of psychophysics and

psychometrics as related to the concept of invariance:

> Psychophysics sees the response as an indicator of an attribute of the individual -- an attribute that varies with the stimulus and is relatively invariant from person to person. Psychometrics regards the response as indicative of an attribute that varies from person to person but is relatively invariant for different stimuli. Both psychophysics and psychometrics make it their business to display the conditions and limits of these invariances.
>
> (Stevens, 1951, p. 31)

The first sentence in this quotation illustrates the idea of sample-invariant item calibration, while the second sentence points to the idea of item-invariant measurement of individuals. This duality between psychophysics and psychometrics, which was clearly described by Mosier (1940, 1941) and pointed out even earlier by Guilford (1936), represents one of the five major ideas underlying test theory identified by Lumsden (1976). Measurement problems related to invariance can be meaningfully viewed in terms of these two broad classes -- sample-invariant item calibration and item-invariant measurement of individuals.

Within each of these two classes, invariance over methods and conditions can be examined. Methods refer to the statistical procedures and models, including the method used to collect the

data, used within the measurement theory. For example, paired comparison and successive interval scaling would represent different methods of data collection, as well as require different statistical models. Conditions can refer to either subgroupings of items and/or examinees. For example, test equating is concerned with the development of procedures which yield comparable estimates of an individual's ability which are invariant over the subgroups of items (tests) which are used to obtain these ability estimates. As another example, the research on item bias or differential item functioning as it has come to be labelled, reflects a concern with whether or not the meaning of an individual's responses on a particular test item vary as a function of irrelevant factors related to membership in various social categories, such as gender, race and social class groups.

Sample-invariant item calibration

The basic measurement problem underlying sample-invariant item calibration is how to minimize the influence of arbitrary samples of individuals on the estimation of item scale values. For example, Engelhard (1984) found that Thorndike provided a single adjustment (location) for differences in group characteristics, while Thurstone provided for two adjustments (location and scale). Rasch's approach to sample-invariant calibration can be viewed as providing three adjustments (location, scale and an individual

level response model). Andrich (1978) has also provided an important comparison between Thurstone and Rasch approaches to item scaling using paired comparison responses which also leads to sample-invariant item calibrations.

The overall goal of sample-invariant calibration of items is to estimate the location of items on a latent variable of interest which will remain unchanged across subgroups of individuals and also across various subgroups of items. For example, if the goal of sample-invariant calibration is achieved, then the item scales values will not be a function of subgroup characteristics, such as ability level, gender, race and social class. Further, the calibration of the items should also be invariant over subsets of items, so that if we are developing a calibrated item bank, the scale values of the items are not affected by the inclusion or exclusion of other items in the bank.

Item-invariant measurement of individuals

Turning now to item-invariant measurement, the basic measurement problem involves minimizing the influence of the particular items which happen to be used to estimate an individual's ability. This problem is also related to the scaling and equating of test scores, as well as the scoring of each individual's performance. Solutions to this problem usually include adjustments for item characteristics (item difficulty) and

test characteristics (location, dispersion and shape of item distributions on the latent variable scale). The overall objective is to obtain comparable estimates of individual ability regardless of which items are included in the test. This is essentially the problem of equating person measurements obtained on tests composed of different items (Engelhard & Osberg, 1983). Invariance over scoring method also requires attention. In addition to considering invariance over methods, it is important to consider invariance over conditions within this context; an individual's score should not depend on the scores of other individuals being tested at the same time.

In summary, invariance can be viewed as an important general concept in the physical and behavioral sciences, as well as a key aspect of successful measurement in the behavioral sciences. As pointed out by Bock and Jones (1968), "in a well-developed science, measurement can be made to yield invariant results over a variety of measurement methods and over a range of experimental conditions for any one method" (p. 9). In outline form, this can be summarized as follows:

Classes of Invariant Measurement

I. Sample-invariant item calibration

A. Invariance over methods (statistical procedures/models)

B. Invariance over conditions (groups of individuals/items)

II. Item-invariant measurement of individuals

A. Invariance over methods (statistical procedures/models)

B. Invariance over conditions (groups of individuals/items)

THREE MEASUREMENT THEORIES AND INVARIANT MEASUREMENT

The purpose of this section is to describe and illustrate how the concept of invariance emerged within the measurement theories of Thorndike, Thurstone, and Rasch. Since the clearest statement of the conditions necessary to accomplish invariance are presented in the measurement theory of Rasch, I will begin with his research and then trace the adumbrations of these ideas within the work of Thurstone and Thorndike. I should also point out that all three of these theorists wrote extensively on various measurement problems, and for Thorndike especially it was sometimes difficult to point to one consistent set of principles that defined his definitive "theory of measurement". In order to address this issue, I have explicitly cited certain texts and it should be understood that I am using these to define a particular individual's "measurement theory". This was not much of a problem for Rasch because he was very consistent in his views related to invariance; Thurstone was fairly consistent, while Thorndike was the least consistent of the three.

Rasch

Based on psychometric research conducted during the 1950s, Rasch (1980/1960, 1961, 1966a, 1966b) presented a set of ideas and methods which were described by Loevinger (1965) as a "truly new approach to psychometric problems" (p. 151) which can lead to "nonarbitrary measures" (p. 151). One of the major characteristics of this "new approach" was Rasch's explicit concern with the development of "individual-centered techniques" as opposed to the group-based measurement models used by measurement theorists such as Thorndike and Thurstone. In Rasch's words, "individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated" (1980/1960, p. xx).

Problems related to invariance play an important role in motivating the measurement theory of Rasch. As pointed out by Andrich (1988a), Rasch presented "two principles of invariance for making comparisons that in an important sense precede, though inevitably lead to, measurement" (p.18). Rasch's concept of "specific objectivity " which he formulated in terms of his principles of comparison form his version of the goals of invariant measurement (Rasch, 1977). In Rasch's words,

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch, 1961, pp. 331-332).

It is clear in this quotation that Rasch recognized the importance of both sample-invariant item calibration and item-invariant measurement of individuals. In fact, he made them the cornerstones of his quest for "specific objectivity". In order to address problems related to invariance, Rasch laid the foundation for the development of a "family of measurement models" which are characterized by separability of item and person parameters. (Masters & Wright, 1984).

Rasch's approach to sample-invariant item calibration involved the comparison of item difficulties obtained in separate groups. In his words,

> In relation to attainment tests all the school grades for which the tests are in practice applicable may be considered

> as forming a total collection of persons, that may be divided into subpopulations, such as single grades, sex groups and age groups within a grade, social strata, etc. Between the test results in such more or less extensive groups the same fundamental relationship must hold, and if so we shall use the term that the relationship is "relatively independent of population", the qualification "relatively" pointing to the degree of breakdown that has been applied to the data.
>
> (Rasch, 1980/1960, p. 9)

In his book, he used ability groups formed on the basis of raw scores. In essence, Rasch was "looking for trouble in a more or less definite direction, namely, for the possibility that the relative difficulties of the tests may vary with [raw score] that is, with the reading inability of the children" (Rasch, 1961, p. 323). This "test of fit" or what Rasch referred to as "control of the model" was presented graphically. In order to illustrate this idea, the results for two subtests, N and F, from the Danish Military Group Intelligence Test (BPP) which were used by Rasch (1980/1960) are presented in Figure 1. The test data were obtained

---

Insert Figure 1 about here

---

from 1,904 recruits who were tested in September 1953. The results

for Subtest N are presented in Panel A (Rasch, 1980/1960, p. 89) which illustrates successful sample-invariant item calibration. The horizontal axis is based on the average of the separate within group calibrations. The parallel lines indicate that the difficulty of the items are relatively invariant across raw-score groups. Unsuccessful sample-invariant item calibrations are presented in Panel B for Subtest F (Rasch, 1980/1960, p. 98) and is reflected in the non-parallel lines with different slopes.

Due to the formal symmetry in Rasch's model between items and individuals, he used a similar graphic approach to examine whether or not item-invariant measurement of individuals had been achieved. The results for Subtests N and F are presented in Figure 2 which

---

Insert Figure 2 about here

---

are also reproduced from Rasch (1980/1960). Panel A (Rasch, 1980/1960, p. 87) illustrates successful item-invariant measurement with ability estimates relatively invariant over item groups, while Panel B (Rasch, 1980/1960, p. 97) provides evidence of unsuccessful item-invariant measurement as evidenced by the inequality of the slopes based on the regression of ability estimates obtained separately within each item group on the total.

Even though there are more sophisticated methods for examining invariance using statistical tests of item and person fit (Wright, 1988; Wright & Stone, 1979), the graphical methods clearly show whether or not invariance has been achieved. As will be seen in the next section, Thurstone used a similar graphical method to examine whether or not his method of absolute scaling was appropriate for a particular set of test data.

By focusing on the individual as the level of analysis, Rasch was able to examine test data and identify when invariance was exhibited. When the data fit the Rasch model, such as with Subtest N, then the types of invariance which eluded research workers in the classical test theory tradition can be obtained. To quote Loevinger,

> Rasch is concerned with a different and more rigorous kind of generalization than Cronbach, Rajaratnam and Gleser. When his model fits, the results are independent of the sample of persons and of the particular items with some broad limits. Within these limits, generality is, one might say, complete.
>
> (Loevinger, 1965, p. 151)

Detailed descriptions of Rasch measurement are presented in Wright and Stone (1979), Wright and Masters (1982) and Wright (1988).

Thurstone

Thurstone also recognized the important of invariant measurement. In fact, as pointed out by Bock and Jones (1968), "in the system of psychological measurement based on the Thurstonian models, we achieve some of the invariance in measurement which is characteristic of the other sciences" (p. 9). In developing his method of absolute scaling (1925, 1927, 1928a, 1928b) for calibrating test items, he was specifically motivated by the lack of sample-invariance he observed in Thorndike's scaling method. In his words,

> the probable error, or PE [used in Thorndike's method], is not valid as a unit of measurement for educational scales. Its defect consists in that it does not possess the one requirement of a unit of measurement, namely constancy [emphasis added]. It fluctuates from one age to another.
>
> (Thurstone, 1927, p. 505)

Thurstone's concept of constancy is his version of an invariance condition and is an explicit consequence of measurement situations that yield objective measurements. Thorndike's PE values fluctuate because the item scale values are not sample-invariant which violates Thurstone's insight that the "scale value of an item should be the same no matter which age group is used in the standardization" (Thurstone, 1928a, p. 119).

As did Rasch, Thurstone used the idea of a continuum to represent the latent variable of interest and assumed that items can be placed at points on this linear scale which would have a fixed position regardless of the group being tested. In order to illustrate this idea, Thurstone presented two figures which are reproduced in Figure 3. The first figure presented in Panel A (Thurstone, 1925, p. 437) shows the location of an item (open

---

Insert Figure 3 about here

---

circle on the base line) which has a fixed position regardless of the distribution of abilities on this latent continuum for groups A and B. According to Thurstone, "if any particular test item or particular raw score is to be allocated on the absolute scale, its scale value should be ideally the same whether determined by group one or group two" (1925, p. 438). In a second figure, shown in Panel B of Figure 3 (Thurstone, 1927, p. 509), shows the location of 7 items (a to g) and again presented the idea that the calibration of these items should be invariant over groups A and B.

In order to adjust for differences in the location and variability of two or more distributions, Thurstone assumed a normal distribution of ability for each group and essentially

adjusted statistically for differences in location and scale. In order for these adjustments proposed by Thurstone to successfully lead to sample-invariant item calibration, Thurstone proposed a graphical test of fit. An example is presented in Panel A of Figure 4 (Thurstone, 1927, p. 513) which shows the plot of the

---

Insert Figure 4 about here

---

item scale values (sigma values) calibrated separately in grades 7 and 8. According to Thurstone,

> If the plot in Fig. 4 [Panel A] should be distinctly non-linear, the present scaling method is not applicable. Non-linearity here shows that the two distributions cannot both be normal on the same scale. If the plot is linear, it proves that both distributions may be assumed to be normal on the same scale or base line. (Thurstone, 1927, p. 513).

This "test of fit" can also be presented in the style of the graphical displays used by Rasch; this is shown in Panel B of Figure 4 (Engelhard, 1984, p. 33) for the same data.

The effects of using Thurstone's method of absolute scaling which provides adjustments for differences in the locations and variations of the ability distributions, as compared to Thorndike's scaling method which simply adjusts for location differences is

shown in Figure 5. In Panel A of Figure 5 (Thurstone, 1927, p. 506), the results of using Thorndike's method to calibrate a

---

Insert Figure 5 about here

---

language scale developed by Trabue (1916) is presented; the average language ability increases as a function of grade level, while the variances remain constant. The results obtained by using Thurstone's method are presented in Panel B of Figure 5 (Thurstone, 1927, p. 515); in this figure, average ability increases with grade level, but the variances of the scores also increase. These results seem theoretically plausible. Thurstone's method of absolute scaling is described and illustrated in detail in Engelhard (1984). An "experimental" adjustment for sample effects which occurs with Thurstone's model for paired comparisons is described in Andrich (1978).

Thurstone's method of absolute scaling can also be used to scale test scores (Gulliksen, 1950), but a more interesting discussion of issues related to item-invariant measurement is presented by Thurstone (1926) in an article on the scoring of individual performance. In this article, Thurstone presented a set of conditions as follows:

1. It should not be required to have the same number of test elements at each step of the scale.

2. It should be possible to omit several test questions at different levels of the scale without affecting the individual score.

3. It should be possible to include in the same scale two forms of test.

4. It should not be required to submit every subject to the whole range of the scale. The starting point and terminal point, being selected by the examiner, should not directly affect the individual score.

5. It should be possible to use the scale so that a rational score may be determined for each individual subject and so that the performance of groups of subjects may be compared.

6. The arithmetical labor in determining individual scores should be a minimum.

7. The procedure should be as far as possible consistent with psychophysical methods so that it will be free from the logical errors involved in the Binet scales and its variants.

Conditions one to five clearly show Thurstone's concern with item-invariant measurement. In his 1926 paper, he goes on to propose a scoring method which meets these conditions; it is beyond the scope of this paper to present Thurstone's approach in detail, however, it appears that he was essentially proposing what would be recognized today as "person characteristic curves".

Many of Thurstone's articles on scaling are included in The measurement of values (1959), although his work on absolute scaling is not included in that volume. The technical details and elaborations of Thurstonian models are presented in Bock and Jones

(1968), and Andrich (1988c) provides a useful overview of Thurstone's contributions to measurement theory. Although it is not directly relevant for this paper, it is interesting to note that Thurstone (1947), as did Rasch (1953), also used the concept of invariance as an important aspect of his approach to factor analysis.

Thorndike

In 1904, Thorndike published the first edition of his highly influential book entitled An Introduction to the Theory of Mental and Social Measurements. Thorndike's major aim in writing this book was to "introduce students to the theory of mental measurements and to provide them with such knowledge and practice as may assist them to follow critically quantitative evidence and argument and to make their own researches exact and logical (1904, p. v). Thorndike's book was the standard reference on statistics and quantitative methods in the mental and social sciences for the first two decades of this century (Clifford, 1984; Engelhard, 1988; Travers, 1983). Much of this influence can be attributed to Thorndike's clear and expository writing style. He explicitly acknowledged that contemporary work in measurement theory had not been presented in a manner suitable for students without fairly advanced mathematical skills, and he set out to present a less mathematical introduction to measurement theory based on the belief

that "there is, happily, nothing in the general principles of modern statistical theory but refined common sense, and little in the techniques resulting from them that general intelligence can not readily master" (p. 2).

Thorndike wrote extensively on educational and psychological measurement, covering topics which ranged from the general statement of his theory (Thorndike, 1904) to the measurement of a variety of educational outcomes (Thorndike, 1910, 1914, 1921), as well as intelligence (Thorndike, et al., 1926).

What were the basic measurement problems identified by Thorndike? Thorndike clearly stated that the "special difficulties" of measurement in the behavioral sciences are

1. Absence or imperfection of units in which to measure
2. Lack of constancy in the facts measured
3. Extreme complexity of the measurements to be made.

In order to illustrate the problems related to the absence of an accepted unit of measurement, Thorndike (1904) pointed out that the spelling tests developed by Joseph Mayer Rice did not have equal units. Rice assumed that all of his spelling words were of equal difficulty, while Thorndike argued that the correct spelling of an easy versus a hard word did not reflect equal amounts of spelling ability. Because the units of measurement are unequal, Thorndike asserted that Rice's results were inaccurate. Without

general agreement on units, the meaning of our test scores become more subjective. Within the framework of this paper, Thorndike was illustrating that obtained scores may not be invariant over subsets of items which vary in difficulty.

Inconstancy is the second major measurement problem identified by Thorndike (1904). Many of the measurement problems encountered in the behavioral sciences are related to random variation inherent in human characteristics. Not only are these variations due to the unreliability of our tests, but they also reflect within subject fluctuations. For example, if we measure a person's motivation, or even body temperature repeatedly, these values tend to vary. Thorndike's concept of "constancy" come closest to the idea of invariance as developed in this paper

The final measurement problem or "special difficulty" identified by Thorndike pertains to the extreme complexity of the variables and constructs that we wish to measure. This problem reflects a concern with dimensionality. Most of the variables worth measuring in the behavioral sciences do not readily translate into unidimensional tests which permit the reporting of a single score to represent the individual's location on the latent variable or construct of interest. As pointed out by Jones and Appelbaum (1989), if unidimensionality is obtained for all items and over all groups of examinees, then item parameters will be

invariant across groups and ability parameters will be invariant across items. Methods for conducting item factor analyses designed to explore this issue have been summarized by Mislevy (1986) and an approach to this problem has been illustrated by Muraki and Engelhard (1985).

Thorndike's method for obtaining sample-invariant item calibration is very similar to Thurstone's method of absolute scaling. As described by Thurstone,

> Thorndike's scaling method consists in first determining the scale value of each item for each grade separately with the mean of each grade as an origin. The difficulty of a test item for Grade V children for example, is determined by the proportion of right answers to the test item in that grade. When a test item has been scaled in several grades, the scale values so obtained will of course be different because of the fact that they are expressed as deviations from different grade means as origins. Thorndike then reduces all these measurements to a common origin in the construction of an educational scale by adding to each scale value the scale value of the mean of the grade (Thurstone, 1927, p. 508).

The major difference between Thorndike's method of item scaling and Thurstone's method of absolute scaling is that Thorndike assumed

that the variances of the groups are equal. Thurstone criticized this assumption,

> . . . it is clear that in order to reduce the overlapping sentences or test items to a common base line or scale it is necessary to make not one but two adjustments. One of these adjustments concerns the means of the several grade groups and this adjustment is made by the Thorndike scaling methods. The second adjustment which is not made by Thorndike concerns the variation in dispersion of the several groups when they are referred to a common scale (Thurstone, 1927, p. 509).

The results of using the two different methods were presented earlier in Figure 5. In his later work, Thorndike did include an adjustment for the range of scores (Thomson, 1940).

An explicit statement of Thorndike's views of item-invariant measurement of individuals was not found. Essentially, Thorndike recommended that tests be constructed with items that are equally spaced in terms of their scale values and that the number of items right be used as a person's score.

COMPARISON AND DISCUSSION OF THREE MEASUREMENT THEORIES

A comparison of the major similarities and differences between the measurement theories of Thorndike, Thurstone and Rasch are summarized in Table 1. These three measurement theorists were all

Insert Table 1 about here

working within a scaling tradition and based many of their proposed methods for calibrating test items and measuring individuals on statistical advances made within the field of psychophysics. One of the differences between psychophysics and psychometrics is that the independent variable is usually an observable variable in psychophysics, while in psychometrics the construct is usually unobservable. Since this construct is not directly observable, these three psychometricians used the idea of a latent continuum to represent this unobservable variable.

Although they all held similar positions on these three issues, there are also several important differences between Thorndike and Thurstone as compared to Rasch. One of the major differences is the recognition by Rasch that measurement models can and should be developed based on the responses of individuals to single test items. This focus on the individual, rather than on groups, allowed Rasch to avoid making unnecessary assumptions regarding the distribution of abilities which were used by both Thorndike and Thurstone. As pointed out earlier, Thorndike's method of scaling test items and Thurstone's method of absolute scaling were both based on the assumption that abilities were

normally distributed. By using the individual and not the group, as the level of analysis, Rasch invented a measurement model which was capable of providing estimates of the location of both items and individuals on a latent variable continuum simultaneously. This also allowed Rasch to develop a probabilistic model rather than a deterministic model for the probability of each individual succeeding on a particular test item as a function of his or her ability; this probabilistic relationship is clearly shown in the familiar S-shaped item characteristic curves. Further, by simultaneously including item calibration and individual measurement within one model, he was able to derive "conditional" estimates of these parameters which provides a framework for determining whether or not invariance has been achieved.

SUMMARY

Progress is as difficult to define within the field of measurement as in any other field of study (Donovan, Laudan & Laudan, 1988; Laudan, 1977). The analysis presented here suggests that Rasch's work provides a theoretical and statistical framework for the practical realization of invariant measurement which was sought by both Thorndike and Thurstone. The simultaneous inclusion of both ability and item difficulty within a probabilistic model defined at the individual level of analysis provided a general framework in which item and person parameters can be estimated

separately. Rasch was able to use recent advances in statistics, such as the concept of sufficiency developed by Fisher (1925), to propose an approach to measurement which provides practical solutions to many testing problems related to invariance.

This paper is part of a larger program of research related to the history and philosophy of measurement theory. The overall purposes of this research are to identify basic measurement problems and to describe how these measurement problems are addressed by major measurement theorists. As pointed out earlier, many of the measurement problems that we face today are not new and through the use of historical and comparative perspectives, we can gain a better understanding of both the measurement problems themselves and the progress which has been made toward the solution of these problems. Some of the perennial measurement problems in the behavioral sciences can be viewed as part of the quest for invariant measurement as described in this paper. Another related concept which was not examined here is unidimensionality. A historical and comparative analysis of this concept and its development within scaling theory along the lines used in this paper would be an important contribution to our knowledge of progress in measurement theory.

This paper has focused on the concept of invariance as it has appeared within the context of measurement theory. Invariance can also be viewed more broadly as the quest for generality in science. If we view science in its simplest form as a series of questions and answers, then invariance addresses the problem of whether or not the answers we find are comparable over groups and methods. The concept of invariance within educational and psychological research can also be expanded to include first, second and higher order invariances. For example, invariances of the first order might deal with mean differences between groups on a variable such as math anxiety. A second order concern might be whether or not the correlations between mathematics achievement and anxiety are invariant over gender, social class and race groups. Higher order invariances might relate to the generalizability of a system of inter-relationships between more than two variables.

There are a number of areas for future research related to the manner in which the concept of invariance appears within other measurement theories that are not within the scaling tradition, but derive from the classical test theory tradition. Some illustrative questions are: How does the work on classical test theory fit into the quest for invariance? Wasn't Spearman really looking for an invariant ranking of individuals regardless of time of administration and instrument used? Can the work of Cronbach and

others on generalizability theory be viewed as an attempt to identify and examine sources of error variance in test scores which are related to the concept of "invariance" in educational and psychological tests as presented here? What about invariance within the framework of two and three parameter item response models? What about Guttman's research on psychometrics? What are the explicit connections of classical measurement concepts, such as reliability and validity, to the concept of invariance as presented in this paper? How does invariance relate to unidimensionality?

In summary, the problem of invariance is of fundamental importance for the development of meaningful measures in education and psychology. Item-invariant estimates of individual abilities and sample-invariant estimates item difficulties are essential in order to realize the advantages of objective measurement. The conditions for objective measurement correspond to the concept of invariance as developed in this paper; the conditions for objective measurement are as follows:

> First, the calibration of measuring instruments must be independent of those objects that happen to be used for the calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for the measuring (Wright, 1968, p. 87).

This paper provides a historical and substantive review of the

problems related to item-invariant measurement, as well as illustrating the progress which has been made toward solving measurement problems related to invariance. Further, this paper contributes to an appreciation of Rasch's accomplishments and the elegance of his approach to problems related to item-invariant measurement. As pointed out by Andrich (1988b), Rasch's achievement did not occur in a "historical vacuum" (p. 13) and this paper illustrates how two major measurement theorists, Thorndike and Thurstone, addressed issues which were eventually resolved by Rasch.

REFERENCES

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement, 2, 449-460.

Andrich, D. (1988a). Rasch models for measurement. Newbury Park, CA: Sage Publications, Inc.

Andrich, D. (1988b). A scientific revolution in social measurement. Paper presented at the annual meeting of the American Educational Research Association in New Orleans.

Andrich, D. (1988c). Thurstone scales. In J. P. Keeves (Ed.), Educational Research, Methodology, and Measurement: An International Handbook. Oxford, England: Pergamon Press.

Bock, R. D. & Jones, L. V. (1968). The measurement and prediction of judgement and choice. San Francisco: Holden-Day.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Cattell, J. K. (1893). Mental measurement. Philosophical Review, 2, 373-380.

Clifford, G.J. (1984). Edward L. Thorndike: The sane positivist. Middleton, CT: Wesleyan University Press. (Originally published 1968).

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of

generalizability of scores and profiles. New York: Wiley.

Donovan, A., Laudan, L. & Laudan, R. (1988). (Eds.). Scrutinizing science: Empirical studies of scientific change. Boston: Kluwer Academic Publishers.

Engelhard, G. (1984). Thorndike, Thurstone and Rasch: A comparison of their methods of scaling psychological tests. Applied Psychological Measurement, 8, 21-38.

Engelhard, G. (1988, April). Thorndike's and Wood's principles of educational measurement: A view from the 1980's. Paper presented at the annual meeting of the American Educational Research Association in New Orleans. (ERIC Document Reproduction Service No. ED 295 961).

Engelhard, G. & Osberg, D. W. (1983). Constructing a test network with a Rasch measurement model. Applied Psychological Measurement, 7, 283-294.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver & Boyd.

Guilford, J. P. (1936). Psychometric methods. New York: Mc-Graw Hill Book Company, Inc.

Gulliksen, H. Theory of mental tests. New York: J. Wiley & Sons.

Jones, L. V. (1960). Some invariant findings under the method of successive intervals. In H. Gulliksen & S. Messick (Eds.),

Psychological scaling: Theory and applications, (pp. 7-20). New York: John Wiley & Sons, Inc.

Jones, L. V. & Appelbaum, M. I. (1989). Psychometric methods. Annual review of psychology, 40, 23-43.

Joreskog, K. G. & Sorbom, D. (1986). LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods. Mooresville, IN: Scientific Software, Inc.

Laudan, L. (1977). Progress and its problems: Toward a theory of scientific change. Berkeley, CA: University of California Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694.

Loevinger, J. (1965). Person and population as psychometric concepts. Psychological Review, 72, 143-155.

Lumsden, J. (1976). Test theory. Annual review of psychology, 27, 251-280.

Masters, G. N. & Wright, B. D. (1984). The essential process in a family of measurement models. Psychometrika, 49, 529-544.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics, 11, 3-31.

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. Psychological Review, 47, 355-366.

Mosier, C. I. (1941). Psychophysics and mental test theory II: The constant process. Psychological Review, 48, 235-249.

Muraki, E. & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. Applied Psychological Measurement, 9, 417-430.

Rasch, G. (1953). On simultaneous factor analysis in several populations. Uppsala Symposium on Psychological Factor Analysis. Nordisk Psykologi's Monograph Series, 3.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, (pp. 321-333). Berkeley, CA: University of California Press.

Rasch, G. (1966a). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. Henry (Eds.), Readings in Mathematical Social Science (pp. 89-107). Chicago: Science Research Associates.

Rasch, G. (1966b). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy, 14, 58-94.

Rasch, G. (1980/1960). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press. [Originally published in 1960 by the Danish Institute for Educational Research].

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Spearman, C. (1904). "General intelligence", objectively determined and measured. American Journal of Psychology, 15, 201-293.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103, 677-680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), Handbook of experimental psychology, (pp. 1-49). New York: Wiley.

Thomson, G. H. (1940). The nature and measurement of the intellect. Teachers College Record, 41, 726-750.

Thorndike, E. L. (1904). An introduction to the theory of mental and social measurements. New York: Teachers College, Columbia University.

Thorndike, E. L. (1910). Handwriting. Teachers College Record, 11, 83-175.

Thorndike, E. L. (1914). The measurement of ability in reading. Teachers College Record, 15, 207-277.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. In Whipple, G. M. (Ed.), The seventeenth yearbook of the national society for the study of education. Part II, The measurement of educational products. Bloomington, IL: Public School Publishing Company.

Thorndike, E. L. (1921). Measurement in education. Teachers College Record, 22, 371-379.

Thorndike, E.L., Bregman, E. O., Cobb, M. V. & Woodyard, E. (1926). The measurement of intelligence. New York: Bureau of Publications, Teachers College, Columbia University.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 15, 433-451.

Thurstone, L. L. (1926). The scoring of individual performance. Journal of Educational Psychology, 17, 446-457.

Thurstone, L. L. (1927). The unit of measurement in educational scales. Journal of Educational Psychology, 18, 505-524.

Thurstone, L. L. (1928a). II. Comment by Professor L. L. Thurstone. Journal of Educational Psychology, 19, 117-124.

Thurstone, L. L. (1928b). Scale construction with weighted observations. Journal of Educational Psychology, 19, 441-453.

Thurstone, L. L. (1947). Multiple-factor analysis: A development and expansion of the vectors of mind. Chicago: The University of Chicago Press.

Thurstone, L. L. (1959). The measurement of values. Chicago: The University of Chicago Press.

Trabue, M. R. (1916). Completion-test language scales. Contributions to Education, No. 77. New York: Columbia University, Teachers College

Travers, R. M. W. (1983). How research has changed American schools: A history from 1840 to the present. Kalamazoo, MI: Mythos Press.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 invitational conference on testing problems. Princeton, NJ: Educational Testing Service.

Wright, B. D. (1988). Rasch measurement models. In J. P. Keeves (Ed.), Educational Research, Methodology, and Measurement: An International Handbook. Oxford, England: Pergamon Press.

Wright, B. D. & Masters, G. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press.

Wright, B. D. & Stone, M. H. (1979). Best test design: Rasch measurement. Chicago: MESA Press.

ACKNOWLEDGEMENTS

Support for this research was provided through a Spencer Fellowship from the National Academy of Education. An earlier version of this paper was presented at the Fifth International Objective Measurement Workshop at the University of California, Berkeley (March, 1989). I would like to thank Judy Monsaas and Larry Ludlow for their helpful comments.

Table 1

Comparison of Thorndike, Thurstone and Rasch on Major Issues

Issue	Thorndike	Thurstone	Rasch
Applied psychophysical methods to address measurement problems (Scaling tradition)	Yes	Yes	Yes
Utilized latent variables	Yes	Yes	Yes
Recognized the importance of invariance	Yes	Yes	Yes
Measurement of individuals and calibration of items addressed simultaneously	No	No	Yes
Developed models for individual responses to test items	No	No	Yes

Figure 1

Rasch's approach to sample-invariant calibration

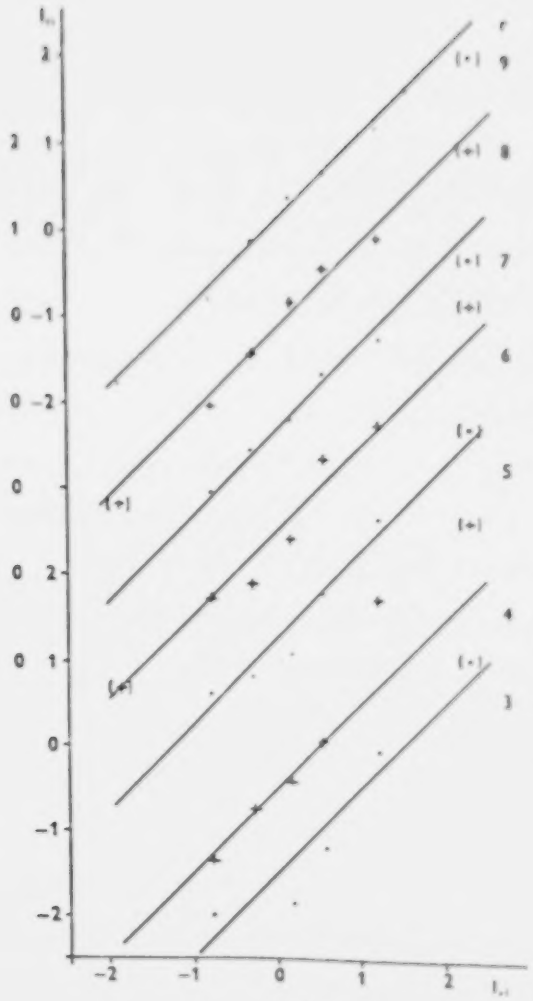


Figure 7
Subtest N of BPP.

A. Successful sample-invariant item calibration

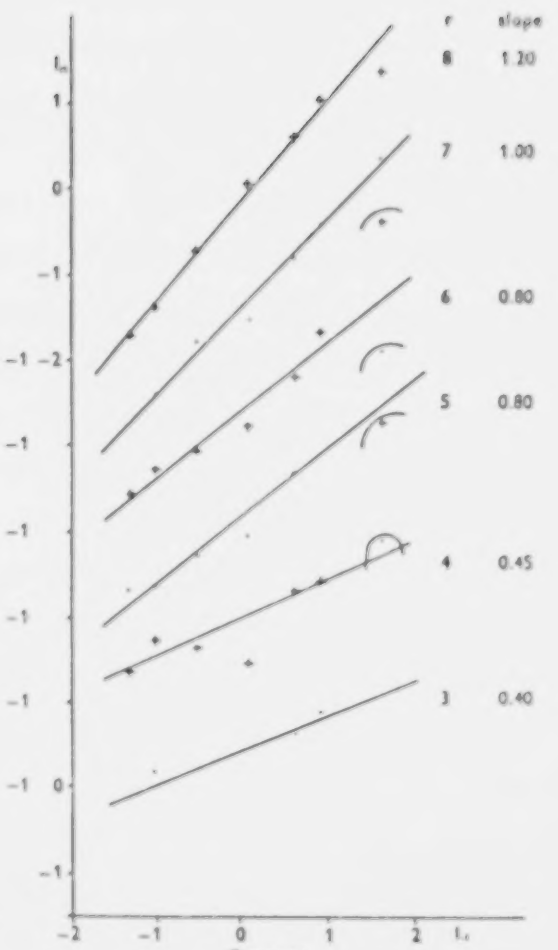


Figure 15
Subtest F of BPP.

B. Unsuccessful sample-invariant item calibration

Figure 2

Rasch's approach to item-invariant measurement

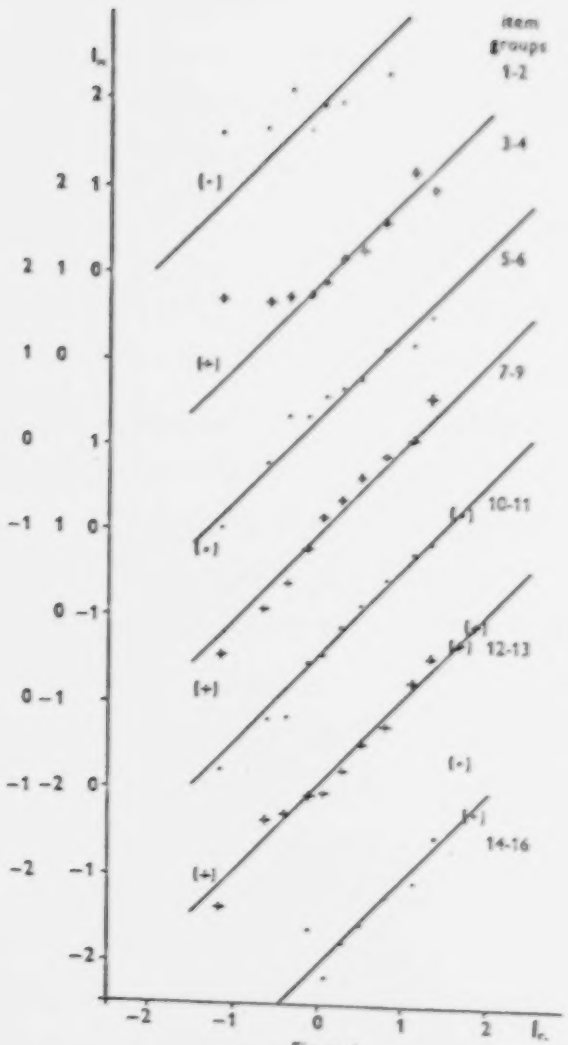


Figure 6
Subtest N of BPP.

A. Successful item-invariant measurement

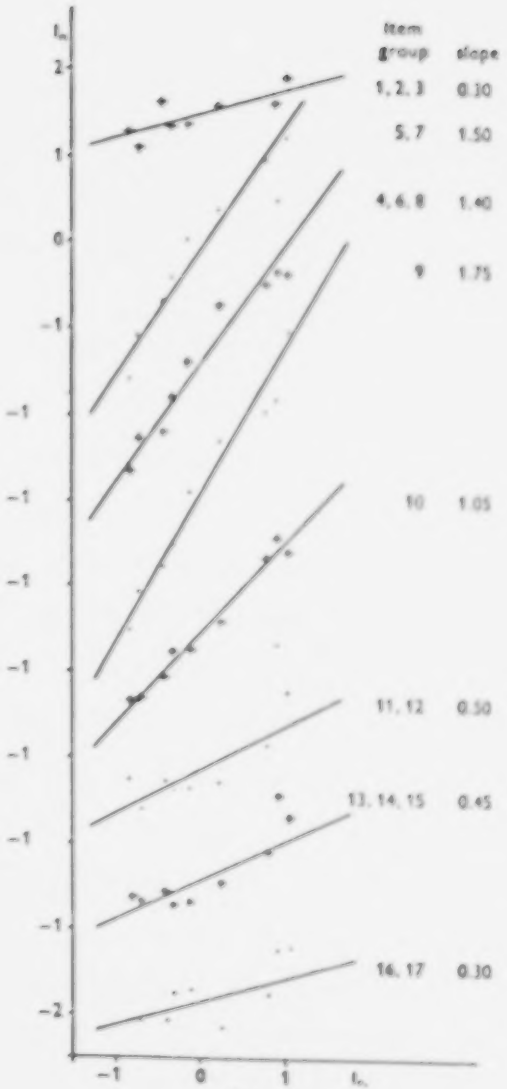
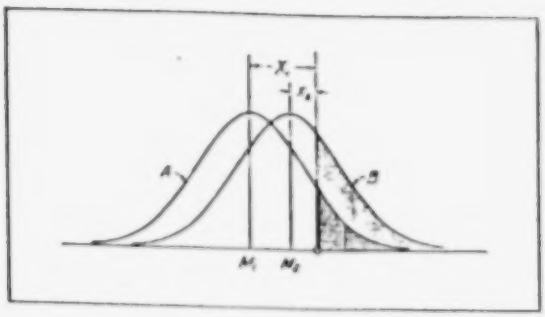


Figure 14
Subtest F of BPP.

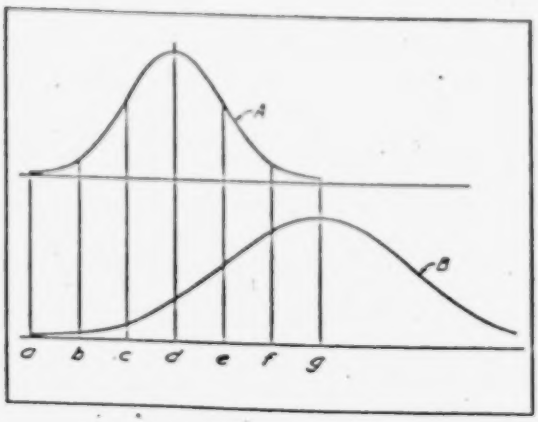
B. Unsuccessful item-invariant measurement

Figure 3

Thurstone's approach to sample-invariant item calibration



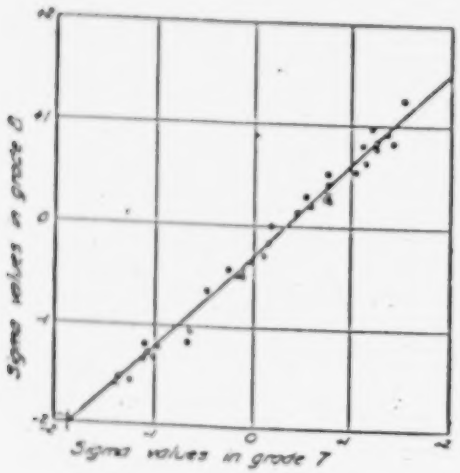
A. Fixed location of one item (open circle) regardless of group (A & B)



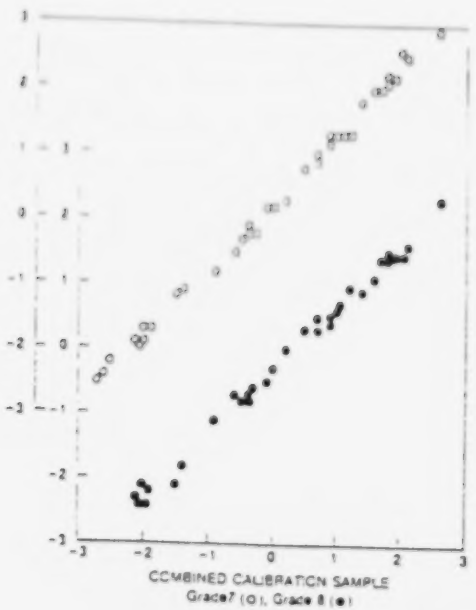
B. Fixed location of seven items (a to g) regardless of group (A & B)

Figure 4

Examining sample-invariant item calibrations



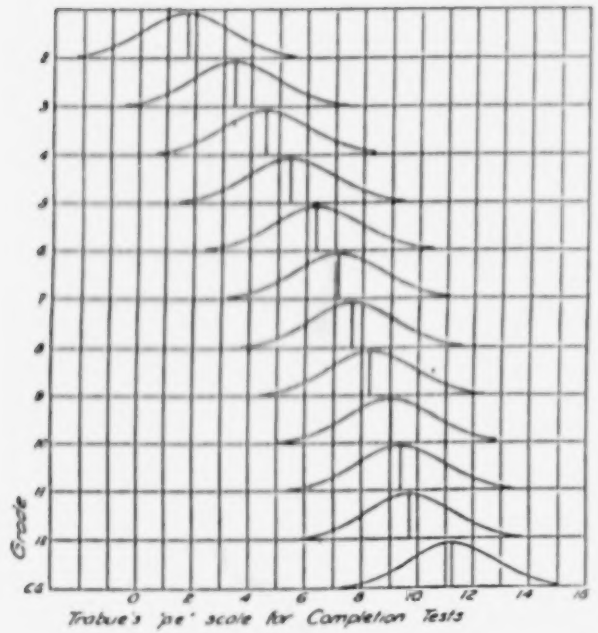
A. Thurstone's test of model fit



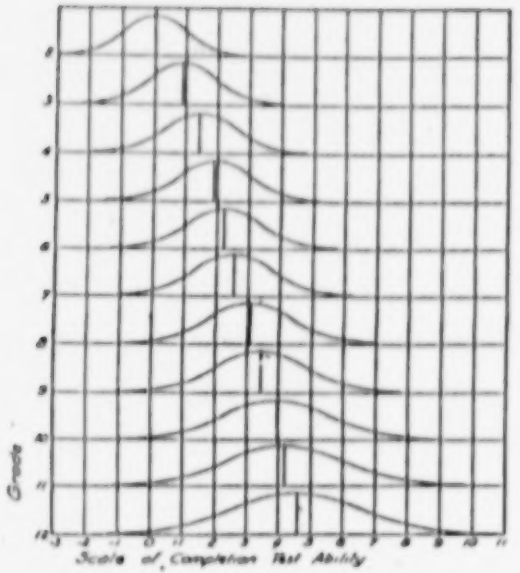
B. Rasch's control of the model

Figure 5

Distribution of language ability in grade 2 to 12 (Trabue's 1916 data)



A. Based on Thorndike's scale



B. Based on Thurstone's method of absolute scaling